REAL TIME ANALYSIS OF STREAMING DATA USING APACHE STORM

Group members:

Prashanth Kurella
Ojasvi Bhalerao
Abhinav Ralhan
Gaurav Malhotra
Anirudh Kumar Rao

Capstone Project

Documentation Report

16.04.2018

Group 34
NU Mentor: Dr. Ram Narayan Yadav, Mr. Ayan Nandy
NIIT Mentor: Mr. Manish Hurkat

Overview

The immense increase in big data applications and analytics has led to creation of specific softwares such as Apache Storm. It is also one of the primary reasons we chose to work on this project, to learn the concepts behind a big data architecture and learn how it works in application.

Here, we implement a live analysis on data and visualizing it. Our aim is to use real time Twitter streaming data for analysis and visualization of big data. Twitter data can be used for various purposes and since it is freely available and also in NoSql format, we chose to work with it. We aim to utilize this data in multiple ways as discussed further. We also create visualizations for all of them. The primary aim, however, is to understand the usage of real time data using Apache Storm and related technologies used in the process.

Goals

1. Sentiment Analysis of Tweets
2. Understanding the Trending topics
3. Geospatial Analysis of data
4. Realizing the reach of tweet

Technologies Used

I. Apache Storm

Apache Storm is an open-source distributed real-time computational system for processing data streams. Similar to what Hadoop does for batch processing, Apache Storm does for unbounded streams of data in a reliable manner.

- Able to process over a million jobs in fraction of a second on a node
- Integrated with Hadoop to harness higher throughputs
- Easy to implement and can be integrated with various programming languages.

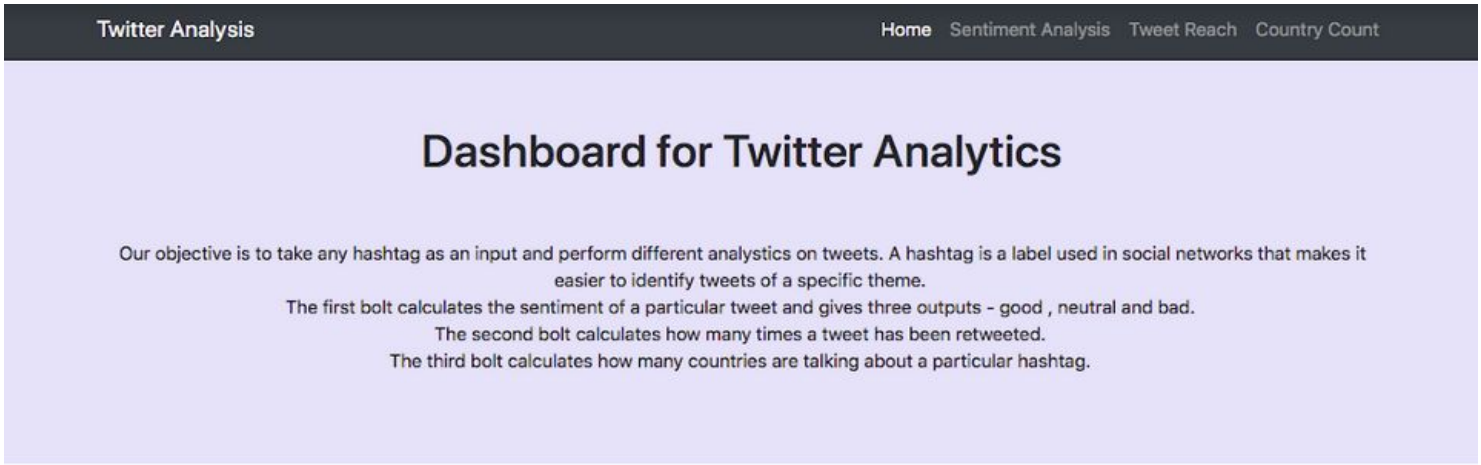
II. MongoDB

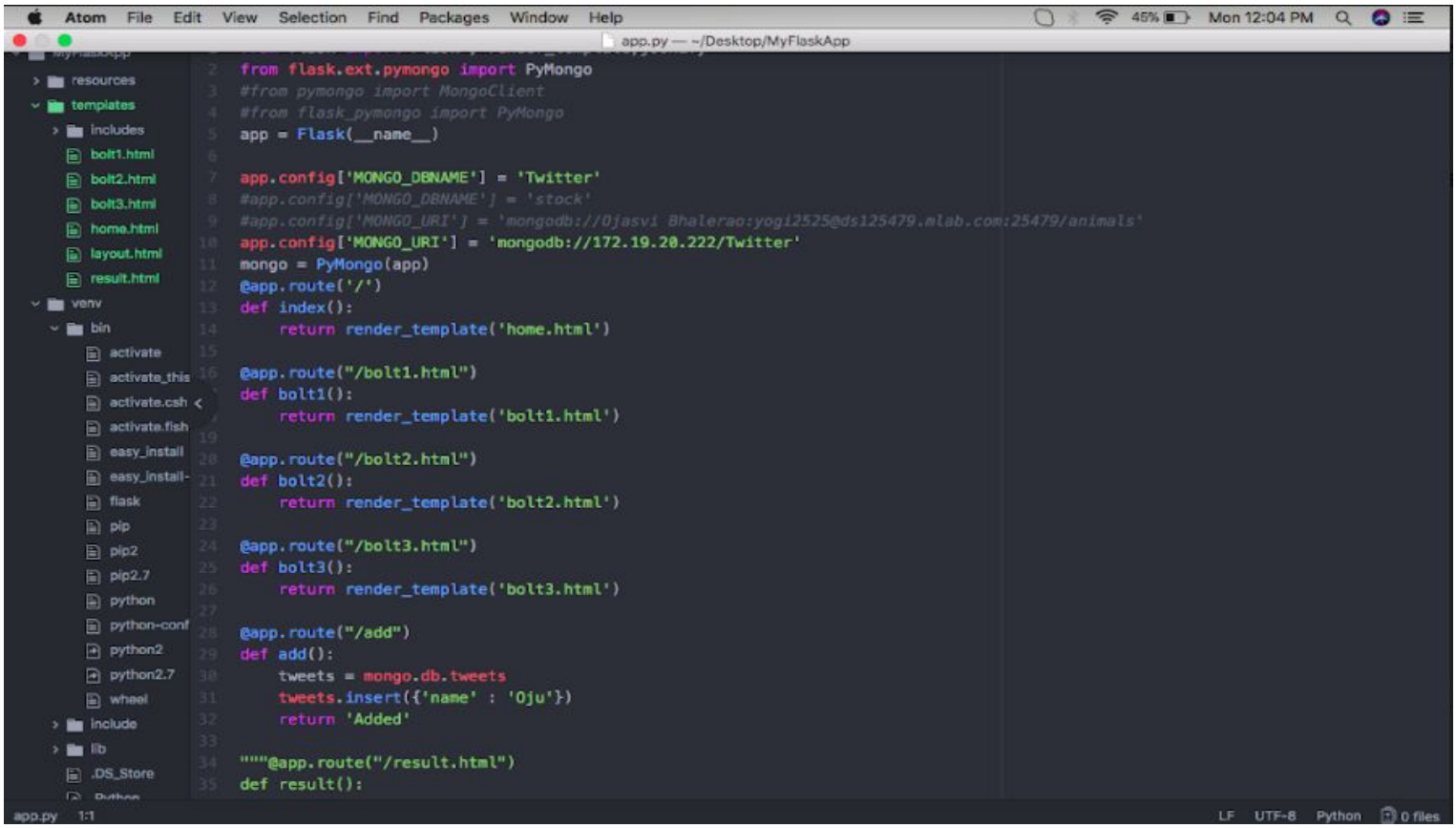
MongoDB is a free and open-source cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemas

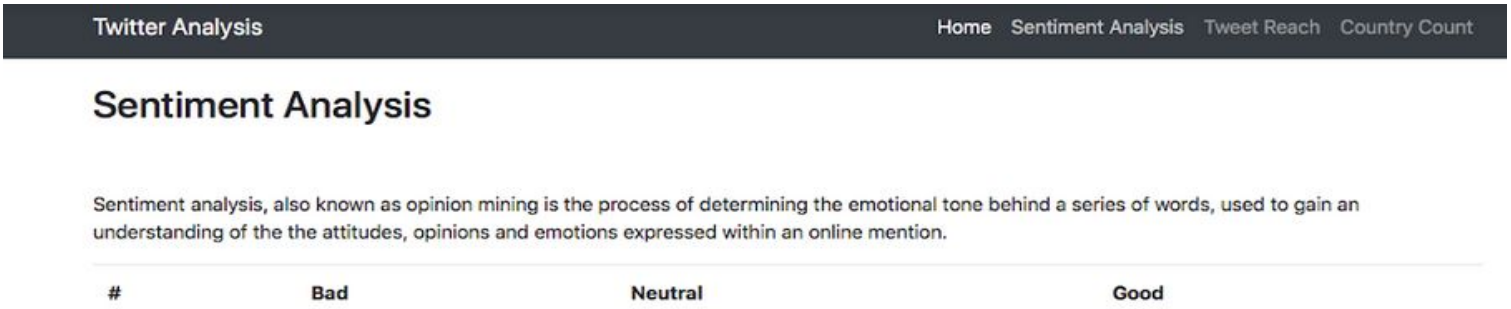
- MongoDB has official drivers for major programming languages and development environments.
- MongoDB supports various features such as Indexing, Replication, Sharding, Aggregation, Load Balancing
- The language drivers are available under the Apache License, and it is available at no cost.

Progress

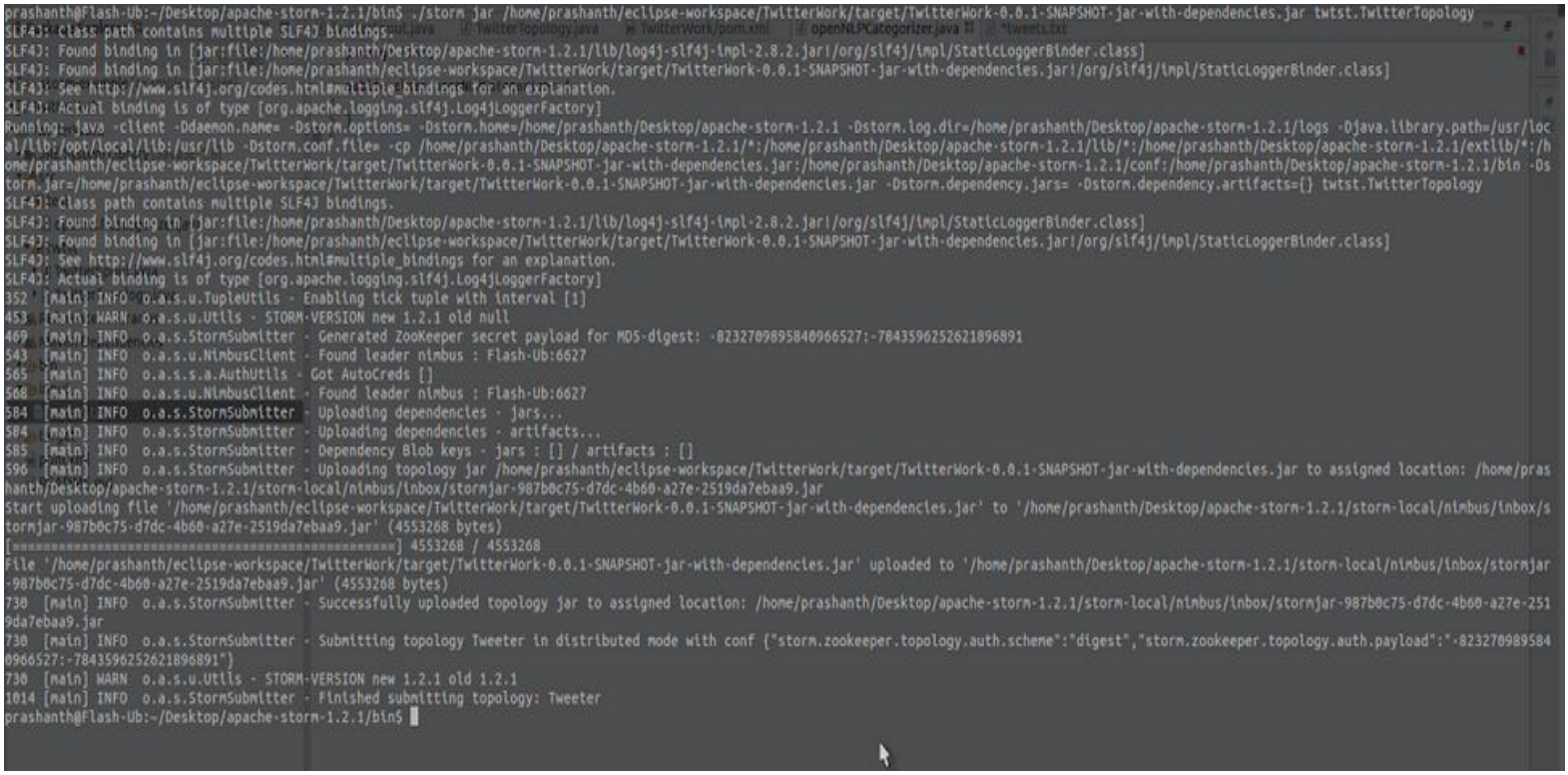
UI to display data



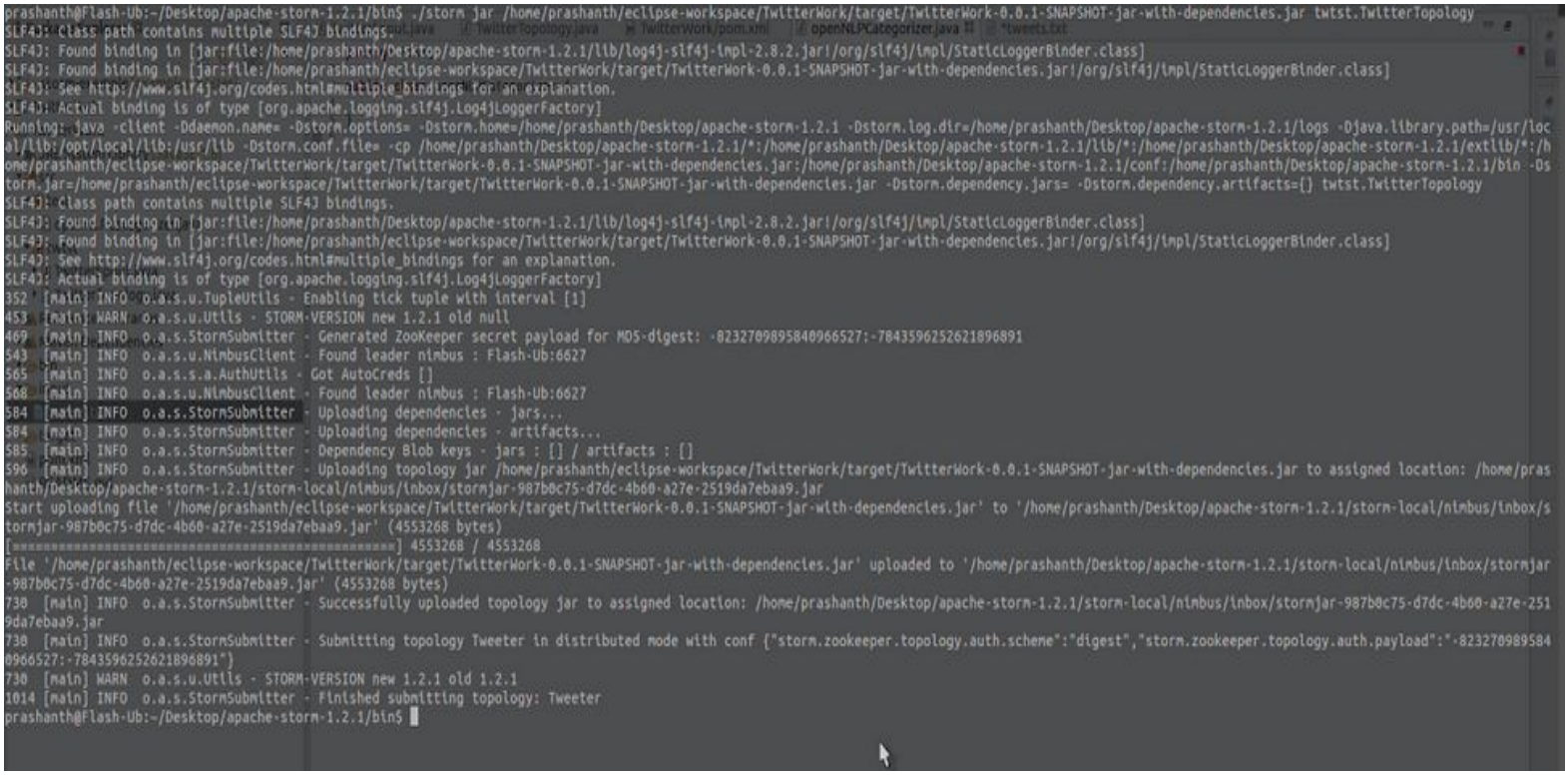




Tweets Retrieved



Submission of Storm Topology



Future Results

1. Write the remaining bolts/spouts and complete our Storm topology.
2. Debug and ensure that data is perfectly generated as per our use cases.
3. To add visualizations for all of our remaining use cases as defined above and add them to the UI.